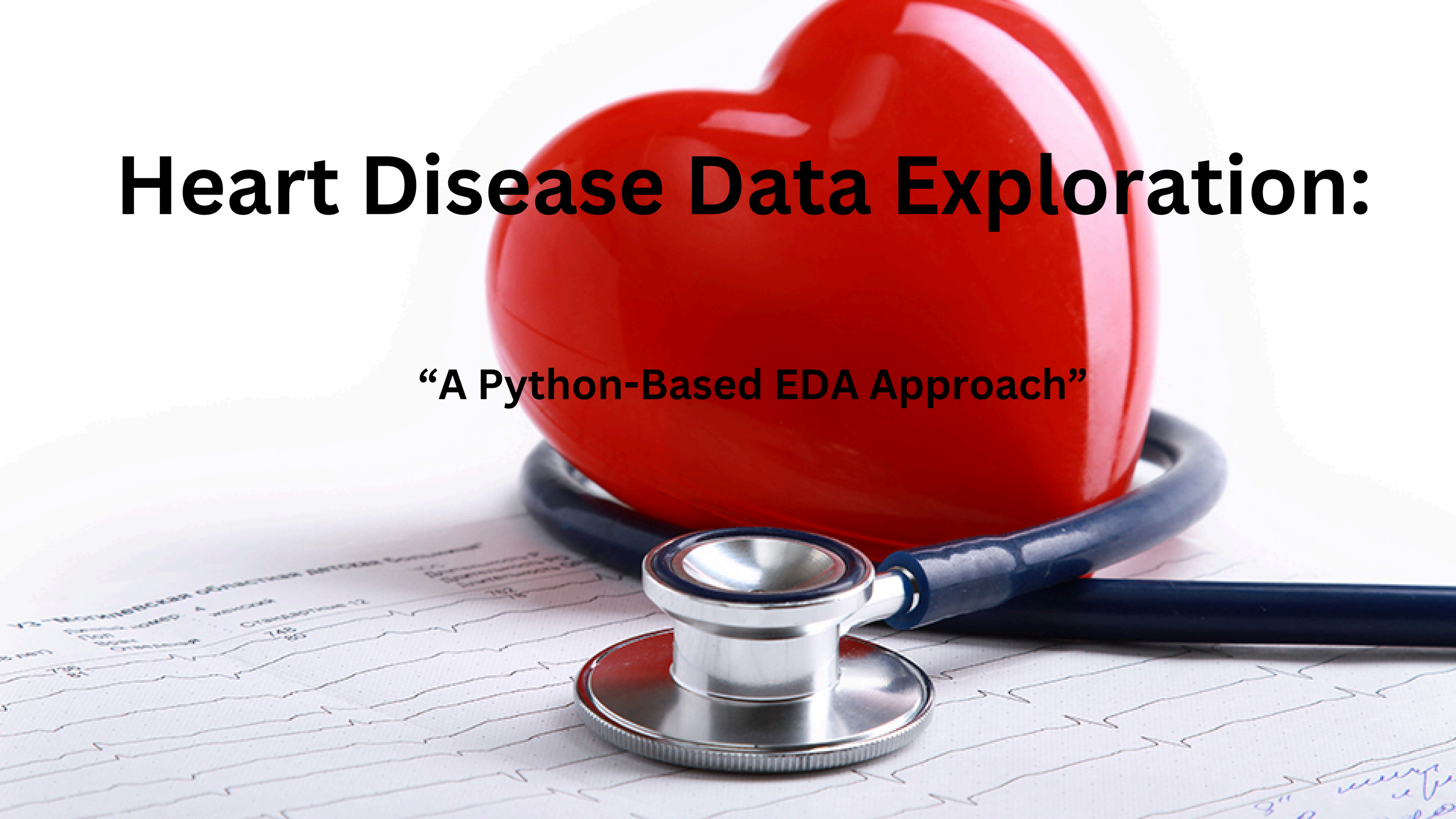
Heart Disease Data Exploration:

“A Python-Based EDA Approach”



EDA

EXPLORATORY DATA ANALYSIS (EDA) IS A CRITICAL STEP IN THE DATA ANALYSIS PROCESS THAT INVOLVES EXAMINING AND SUMMARIZING THE MAIN CHARACTERISTICS OF A DATASET.

IT MAJORLY CONTAINS 5 STEPS :

- DATA CLEANING
- DESCRIPTIVE STATISTICS
- DATA VISUALIZATION
- DATA TRANSFORMATION
- IDENTIFYING RELATIONSHIPS

Let’s start performing EDA

DATA READING:

DATA READING

loading the data

```
[2]  data=pd.read_csv("framingham.csv",header=0)
```

Python

Reading the five rows of data

```
[3]  data.head()
```

Python

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	1	39	4.0	0	0.0	0.0	0	0	0	195.0	106.0	70.0	26.97	80.0	77.0	0
1	0	46	2.0	0	0.0	0.0	0	0	0	250.0	121.0	81.0	28.73	95.0	76.0	0
2	1	48	1.0	1	20.0	0.0	0	0	0	245.0	127.5	80.0	25.34	75.0	70.0	0
3	0	61	3.0	1	30.0	0.0	0	1	0	225.0	150.0	95.0	28.58	65.0	103.0	1
4	0	46	3.0	1	23.0	0.0	0	0	0	285.0	130.0	84.0	23.10	85.0	85.0	0

+ Code + Markdown

DIMENSION OF DATA

MEANS THE NUMBER OF ROWS AND COLUMNS PRESENT IN THE DATA

HERE ARE 4238 ROWS AND 16 COLUMNS

Dimensions of data i.e rows and columns

```
#dimension of data
data.shape
```

```
(4238, 16)
```

Explain the dataset

1. male:

- Type: Binary (0 or 1)

- Description: Indicates the gender of the individual. `1` typically represents male, and `0` represents female.

2. age:

- Type: Numeric (integer)

- Description: The age of the individual in years.

3. education:

- Type: Categorical (ordinal)

- Description: Represents the education level of the individual. This is often coded as numbers representing different levels of education (e.g., 1 for less than high school, 2 for high school, 3 for some college, 4 for college graduate).

4. currentSmoker:

- Type: Binary (0 or 1)

- Description: Indicates whether the individual is a current smoker. `1` means the individual is currently smoking, while `0` means they are not.

5. cigsPerDay:

- Type: Numeric (integer)

- Description: The number of cigarettes the individual smokes per day. Relevant only if `currentSmoker` is `1`.

6. BPMeds:

- Type: Binary (0 or 1)

- Description: Indicates whether the individual is on blood pressure medication. `1` means they are taking blood pressure medication, and `0` means they are not.

7. prevalentStroke:

- Type: Binary (0 or 1)

- Description: Indicates whether the individual has a history of stroke. `1` means they have had a stroke, and `0` means they have not.

8. prevalentHyp:

- Type: Binary (0 or 1)

- Description: Indicates whether the individual has hypertension (high blood pressure). `1` means they have hypertension, and `0` means they do not.

9. diabetes:

- Type: Binary (0 or 1)

- **Description**: Indicates whether the individual has diabetes. `1` means they have diabetes, and `0` means they do not.

10. totChol:

- Type: Numeric (continuous)

- Description: The total cholesterol level of the individual, typically measured in milligrams per deciliter (mg/dL).

11. sysBP:

- Type: Numeric (continuous)

- Description: The systolic blood pressure of the individual, measured in millimeters of mercury (mm Hg). Systolic pressure is the pressure in the arteries when the heart beats.

12. diaBP:

- Type: Numeric (continuous)

- Description: The diastolic blood pressure of the individual, measured in millimeters of mercury (mm Hg). Diastolic pressure is the pressure in the arteries when the heart is at rest between beats.

13. BMI:

- Type: Numeric (continuous)

- Description: Body Mass Index, a measure of body fat based on height and weight. It is calculated as weight in kilograms divided by the square of height in meters (kg/m^2).

14. HeartRate:

- Type: Numeric (continuous)

- Description: The heart rate of the individual, typically measured in beats per minute (BPM).

.

15. glucose:
- Type: Numeric (continuous)
- Description: The blood glucose level of the individual, typically measured in milligrams per deciliter (mg/dL). High levels can indicate diabetes or prediabetes.
16. TenYearCHD:
- Type: Binary (0 or 1)
- Description: Indicates the risk of Coronary Heart Disease (CHD) within the next ten years. `1` means the individual is at risk, and `0` means they are not

Checking the Null values in the dataset:

Upon analysis, it has been determined that the 'Education' feature contains 105 missing values, 'Cigarettes Per Day' has 29 missing values, 'BPMeds' has 53 missing values, 'Total Cholesterol' has 50 missing values, 'BMI' has 19 missing values, 'Heart Rate' has 1 missing value, and 'Glucose' has 388 missing values.

```
[6]  data.isnull().sum()

...  male               0
     age                0
     education        105
     currentSmoker      0
     cigsPerDay        29
     BPMeds            53
     prevalentStroke    0
     prevalentHyp       0
     diabetes           0
     totChol           50
     sysBP              0
     diaBP              0
     BMI               19
     heartRate          1
     glucose          388
     TenYearCHD         0
     dtype: int64
```

Describing the data (Descriptive Analysis)

```
[8]  data.describe()
```

Python

	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP
count	4238.000000	4238.000000	4133.000000	4238.000000	4209.000000	4185.000000	4238.000000	4238.000000	4238.000000	4188.000000	4238.000000	4238.000000
mean	0.429212	49.584946	1.978950	0.494101	9.003089	0.029630	0.005899	0.310524	0.025720	236.721585	132.352407	82.893464
std	0.495022	8.572160	1.019791	0.500024	11.920094	0.169584	0.076587	0.462763	0.158316	44.590334	22.038097	11.910850
min	0.000000	32.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	107.000000	83.500000	48.000000
25%	0.000000	42.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	206.000000	117.000000	75.000000
50%	0.000000	49.000000	2.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	234.000000	128.000000	82.000000
75%	1.000000	56.000000	3.000000	1.000000	20.000000	0.000000	0.000000	1.000000	0.000000	263.000000	144.000000	89.875000
max	1.000000	70.000000	4.000000	1.000000	70.000000	1.000000	1.000000	1.000000	1.000000	696.000000	295.000000	142.500000

1. The mean age of individuals experiencing heart attacks is 49 years, with the youngest recorded case being at 32 years, suggesting a propensity for cardiac events in middle-aged individuals.

2. The average cholesterol level among the studied population is 236.7 mg/dL, significantly exceeding the recommended average level of 200 mg/dL, indicating a prevalent risk factor for cardiovascular diseases.

3. The maximum recorded systolic blood pressure (BP) stands at 295 mmHg, indicative of severe hypertension, a condition associated with heightened susceptibility to heart attacks.

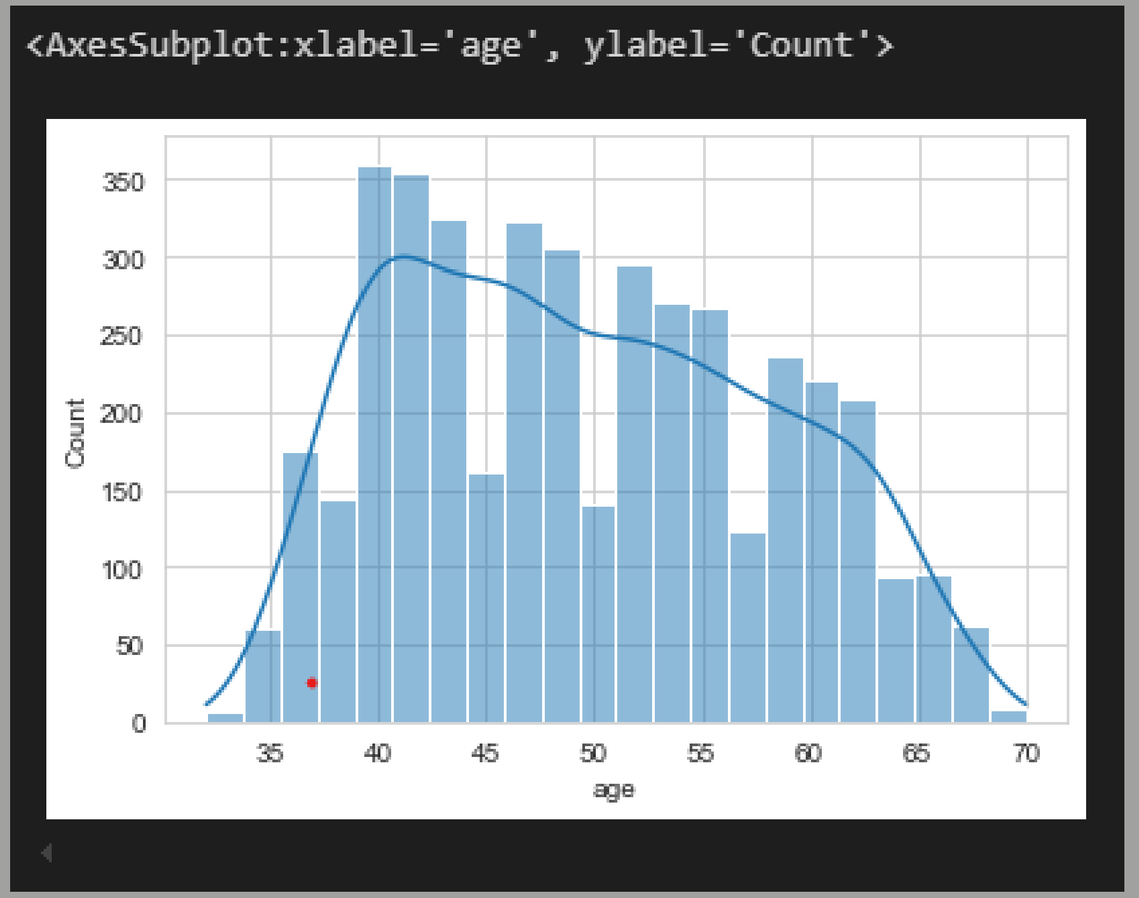
4. The average heart rate among individuals is 75 beats per minute, falling within the normal range of 60 to 100 beats per minute, suggesting overall cardiac health within acceptable parameters.

5. On average, individuals within the cohort consume 9 cigarettes per day, highlighting a prevalent risk factor for cardiovascular diseases, including heart attacks, attributable to smoking habits.

GRAPHICAL EDA

UNIVARIATE ANALYSIS:

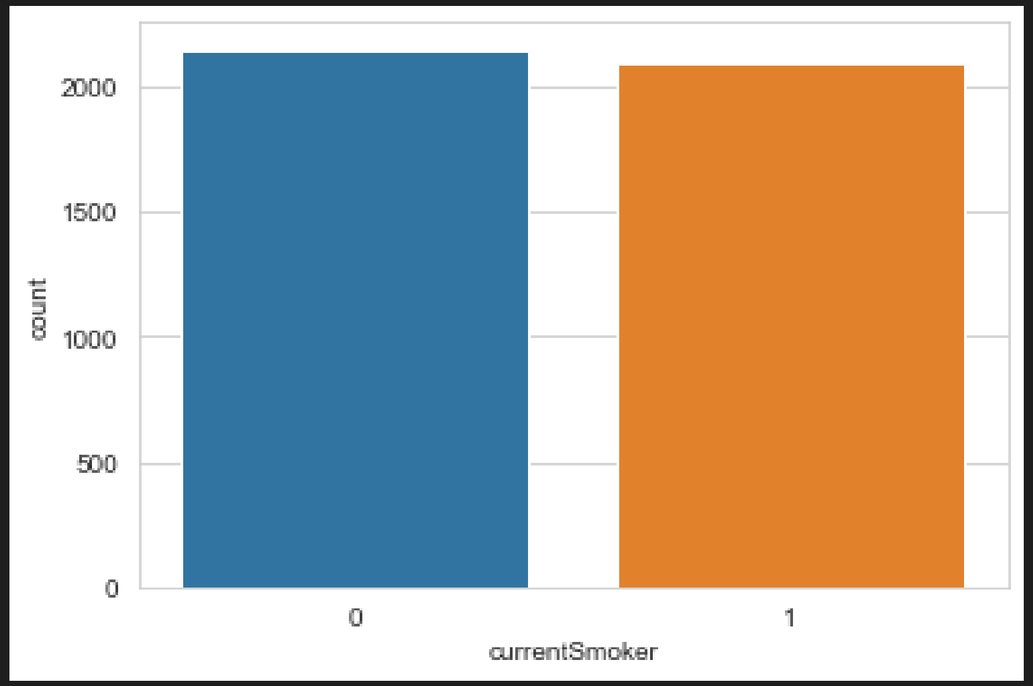
In the analyzed population, age demonstrates a distribution that approximates normality, with a notable concentration observed within the range of 40 to 55 years. Within this age bracket, individuals exhibit an increased susceptibility to experiencing cardiac events, such as heart attacks.



The dataset exhibits an equal distribution, with half of the individuals identified as current smokers and the remaining half as non-smokers.

```
sns.countplot(data=data, x='currentSmoker')
```

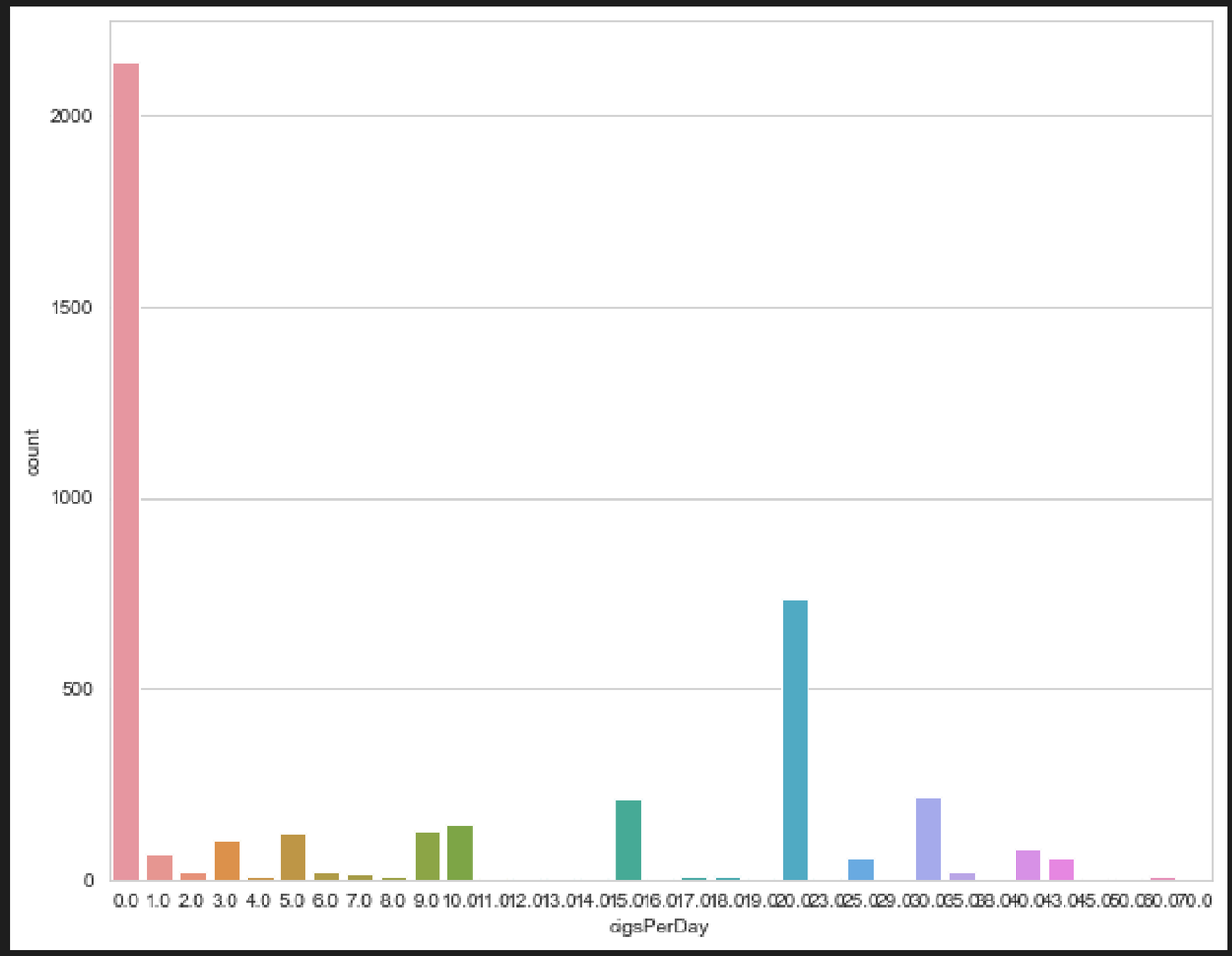
```
<AxesSubplot:xlabel='currentSmoker', ylabel='count'>
```



Although the dataset indicates that the maximum number of cigarettes smoked by individuals is 20, it is noteworthy that the majority of individuals in the sample do not smoke.

[12]

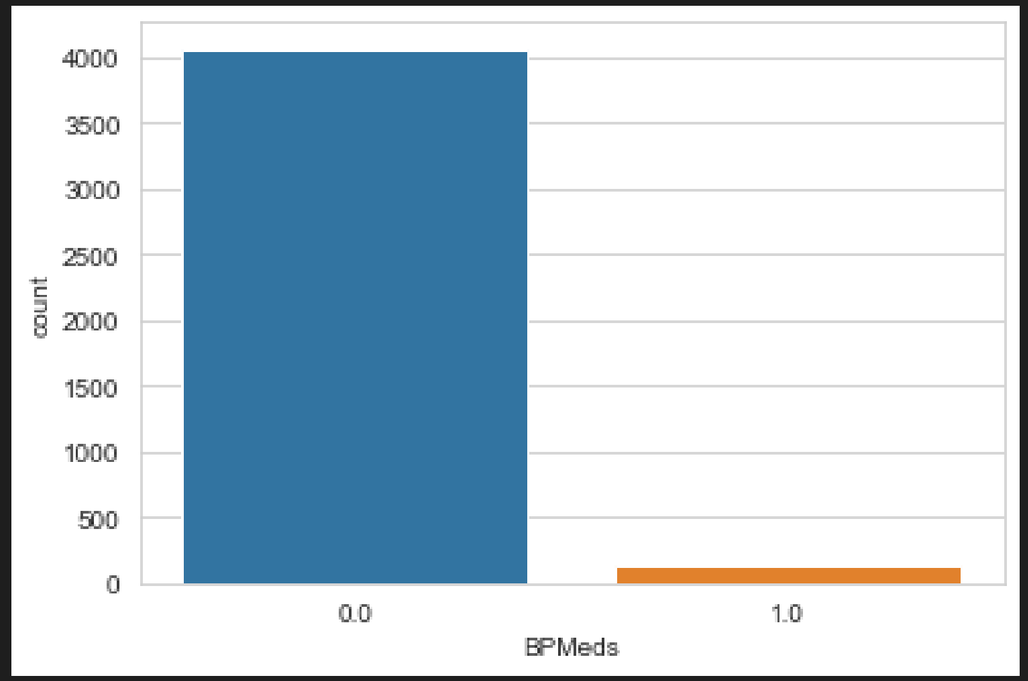
```
<AxesSubplot:xlabel='cigsPerDay', ylabel='count'>
```



The analysis reveals that a significant proportion of individuals within the dataset do not utilize medication for managing blood pressure, while a minority of individuals do.

```
sns.countplot(data=data, x='BPMeds')
```

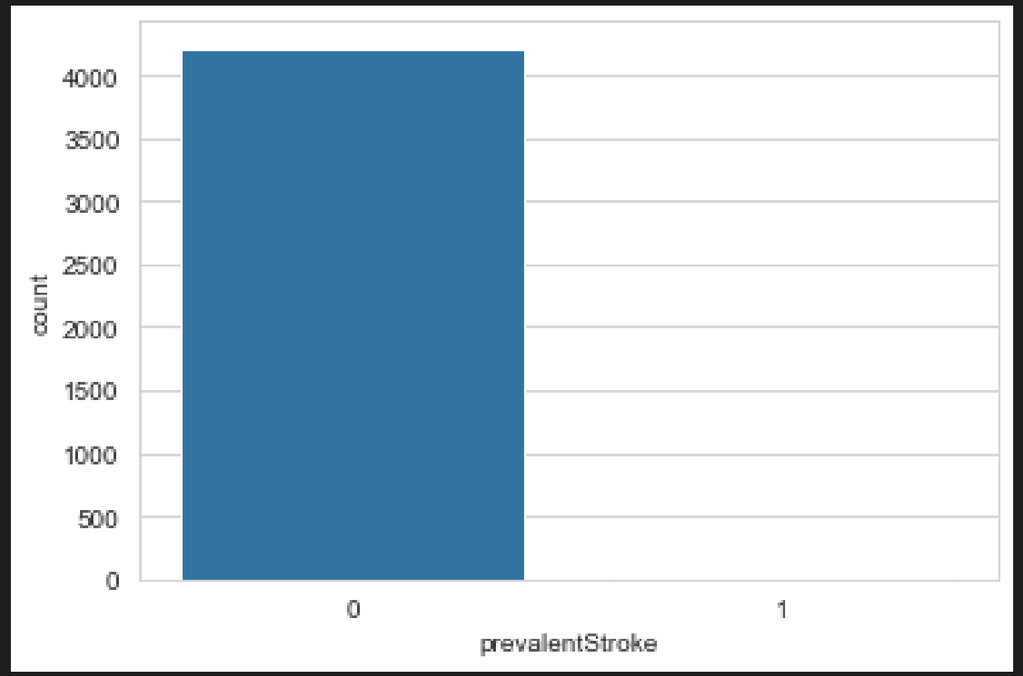
```
<AxesSubplot:xlabel='BPMeds', ylabel='count'>
```



The majority of individuals in the dataset have not experienced a previous heart attack.

```
sns.countplot(data=data, x='prevalentStroke')
```

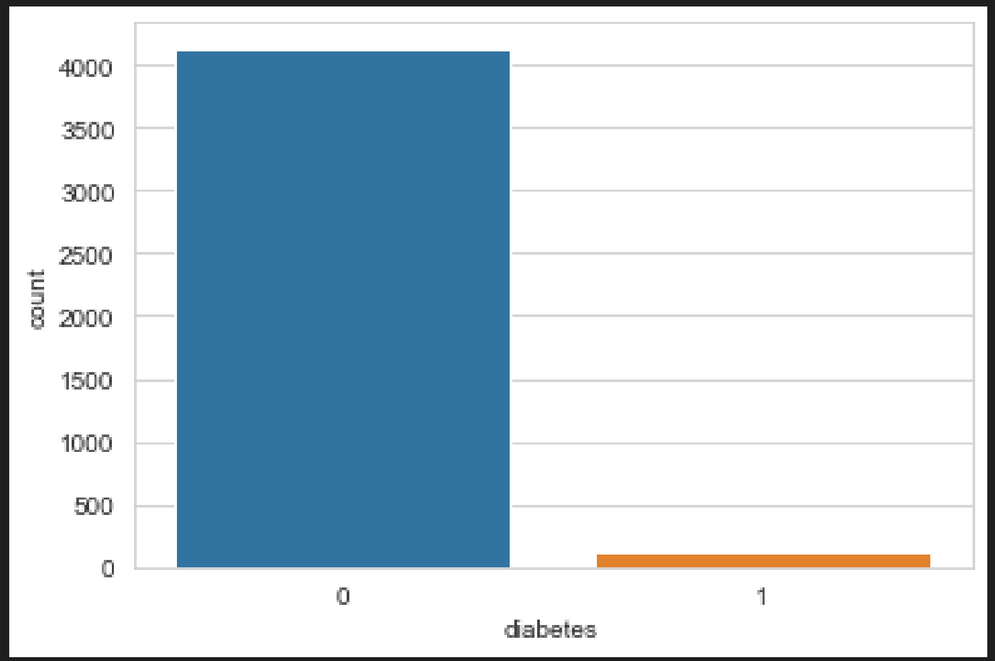
```
<AxesSubplot:xlabel='prevalentStroke', ylabel='count'>
```



The prevailing trend in the dataset indicates that the majority of individuals do not have diabetes.

```
sns.countplot(data=data, x='diabetes')
```

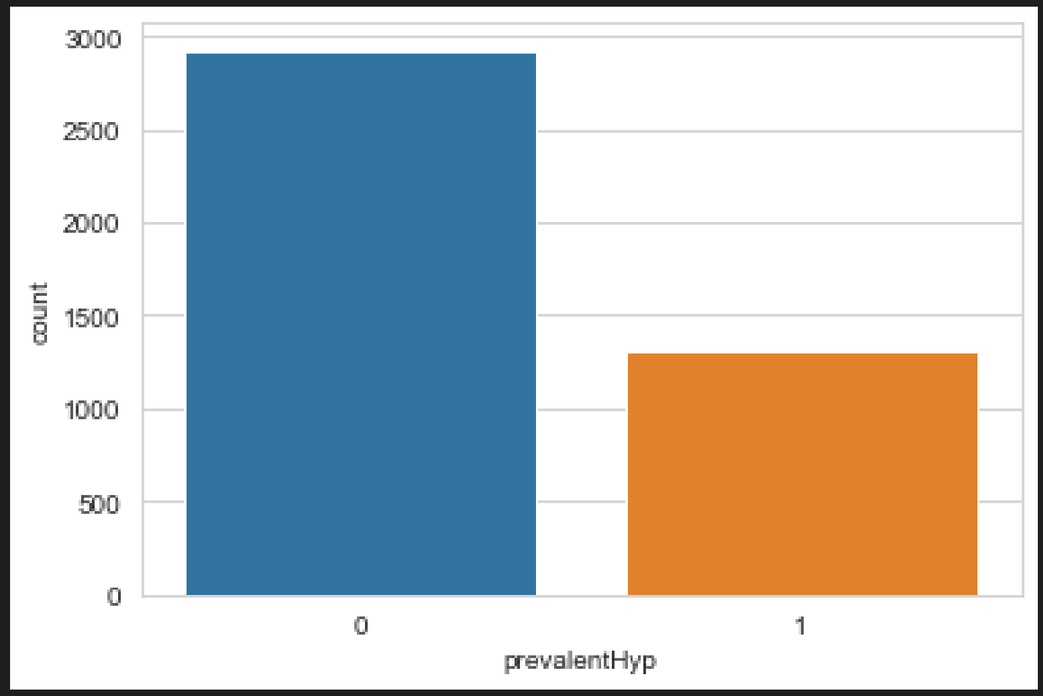
```
<AxesSubplot:xlabel='diabetes', ylabel='count'>
```



A subset of individuals in the dataset has a history of hypertension, a condition that may elevate the risk of experiencing a heart attack.

```
sns.countplot(data=data, x='prevalentHyp')
```

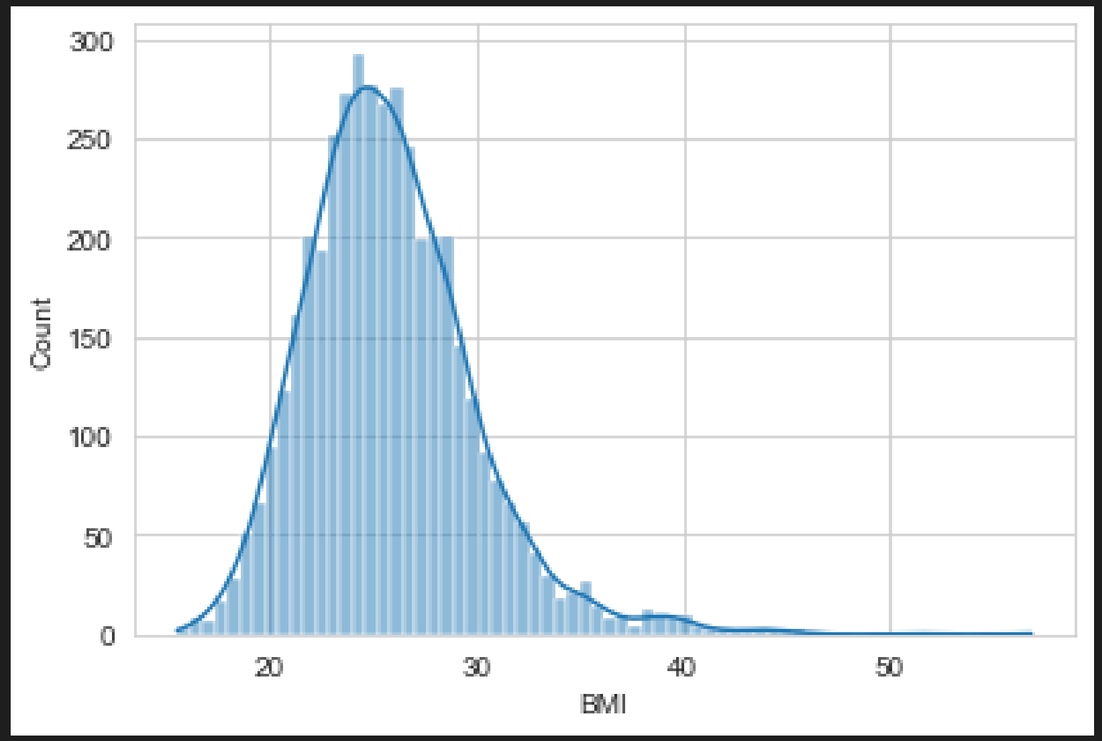
```
<AxesSubplot:xlabel='prevalentHyp', ylabel='count'>
```



The distribution of Body Mass Index (BMI) within the dataset conforms to a normal distribution, with the majority of individuals exhibiting BMI values falling within the range of 20 to 30. This range typically encompasses individuals considered to have a healthy weight or who are moderately overweight

```
sns.histplot(data=data,x='BMI',kde=True)
```

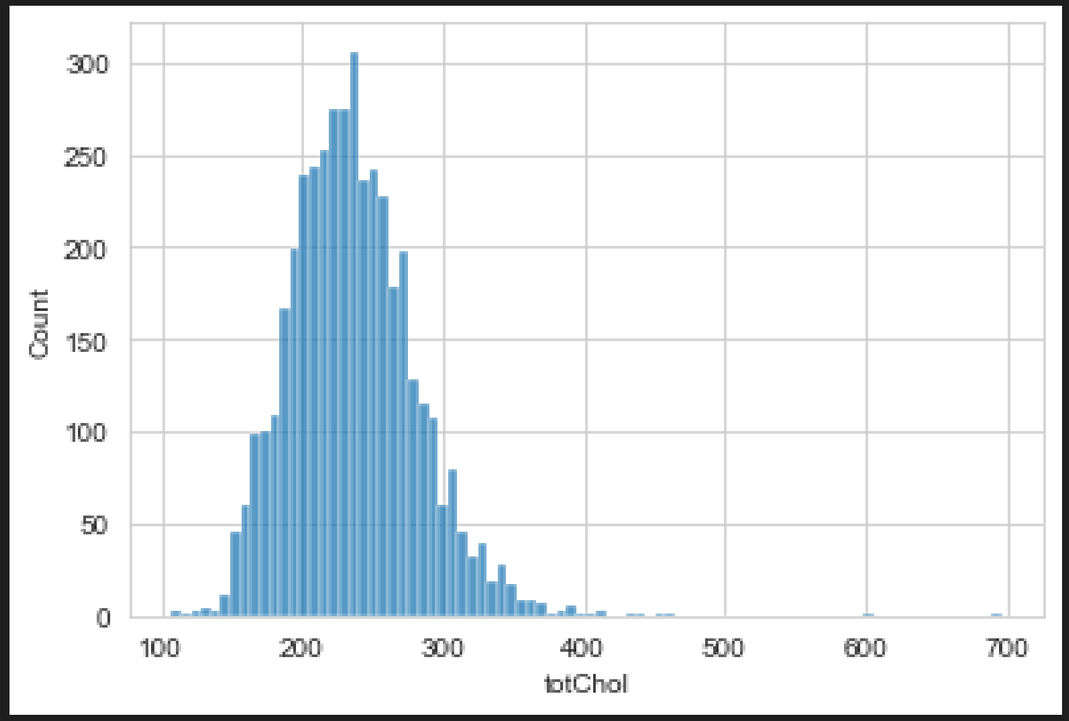
```
<AxesSubplot:xlabel='BMI', ylabel='Count'>
```



The dataset exhibits a slightly peaked distribution for Total Cholesterol levels, with the majority falling between 200 and 300. This distribution hints at varying cholesterol levels among individuals, prompting consideration of potential implications for cardiovascular health and related risk factors.

```
sns.histplot(data=data, x = 'totChol')
```

```
<AxesSubplot:xlabel='totChol', ylabel='Count'>
```



BIVARIATE ANALYSIS

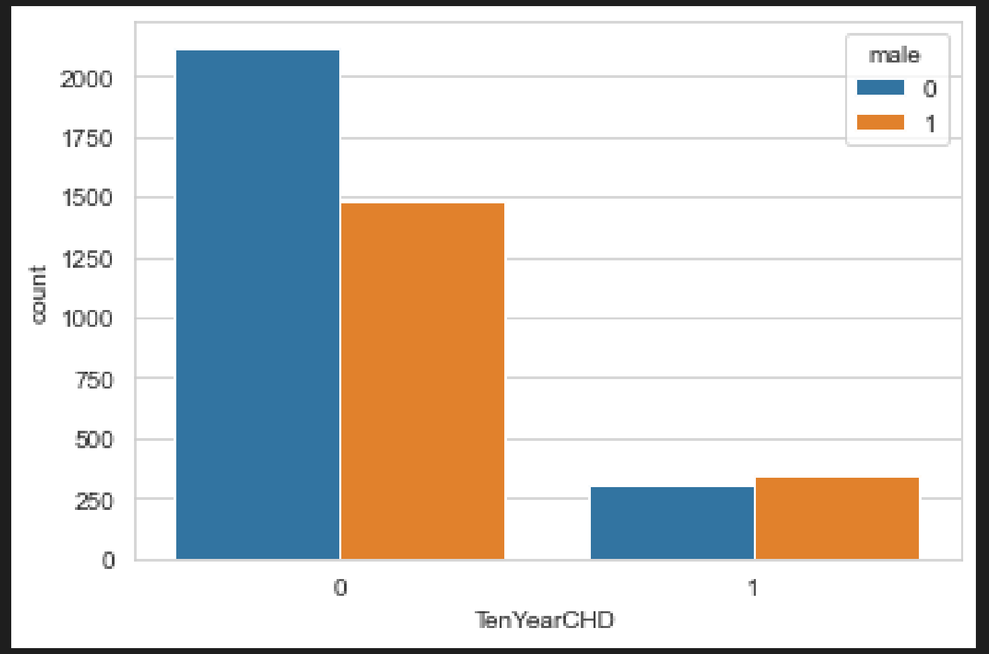
After visualizing the relationship between the 'Male' column and the target variable ('TenYearCHD'), the following insights emerge:

1) Males exhibit a higher susceptibility to experiencing heart attacks compared to females.

2) Females demonstrate a higher likelihood of survival, indicating potentially lower risk or better prognosis regarding coronary heart disease (CHD).

```
#this show that in male  heart disease is high than fema
sns.countplot(x='TenYearCHD',hue='male',data=data)
```

```
<AxesSubplot:xlabel='TenYearCHD', ylabel='count'>
```



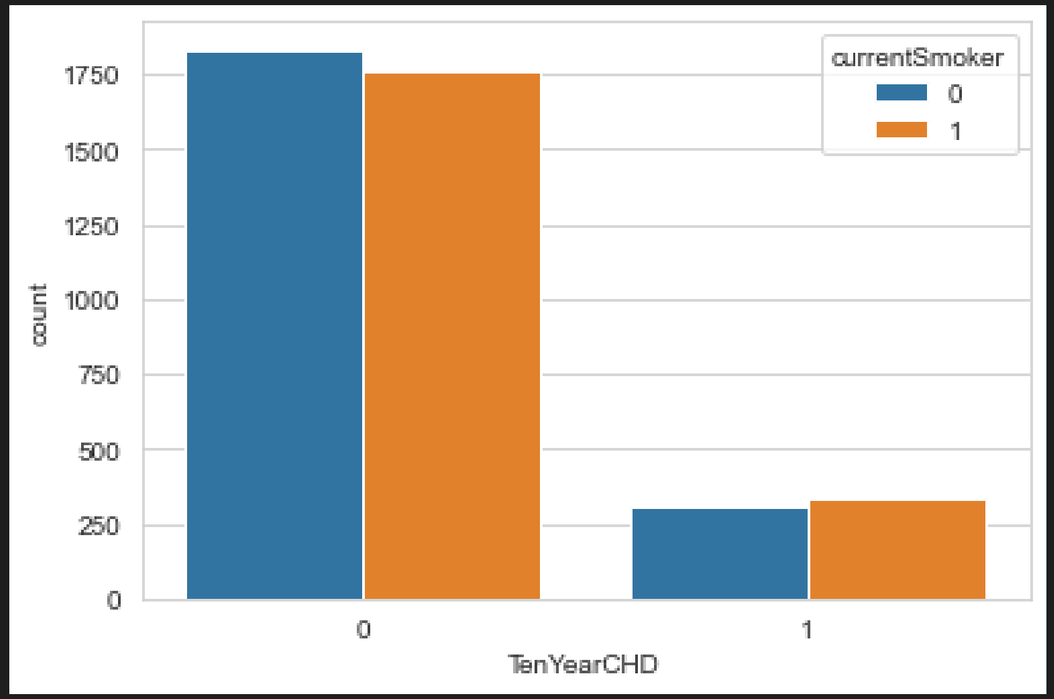
After analyzing the visualized data, the following conclusions can be drawn:

1) Smokers show a higher susceptibility to heart attacks compared to non-smokers.

2) Individuals who smoke face an elevated risk of mortality, indicating a potential correlation between smoking habits and adverse health outcomes.

```
#effect of smoking on ten year CHD
sns.countplot(x='TenYearCHD',hue='currentSmoker',data=data)
```

```
<AxesSubplot:xlabel='TenYearCHD', ylabel='count'>
```



```python
grouped_df = data.groupby('male')[['age','diabetes']].mean()

# Displaying the result
print(grouped_df)
```

```
            age  diabetes
male
0     49.800331  0.023563
1     49.298516  0.028587
```

1)Age Distribution:

-Females (non-males) have an average age of approximately 49.80 years.

-Males have a slightly lower average age of around 49.30 years.

2)Diabetes Prevalence:

-Among females, the prevalence of diabetes is approximately 2.36%.

-Males exhibit a slightly higher prevalence of diabetes, approximately 2.86%.

```
[22]
    grouped_df = data.groupby('TenYearCHD')['diabetes'].count()

    # Displaying the result
    print(grouped_df)
```

```
...  TenYearCHD
     0    3594
     1     644
     Name: diabetes, dtype: int64
```

1)Distribution by Coronary Heart Disease (CHD) Status:

-Among individuals without CHD (TenYearCHD = 0), there are 3,594 cases.

-For individuals with CHD (TenYearCHD = 1), there are 644 cases.

2)Imbalance in CHD Status:

-The dataset exhibits a significant class imbalance, with a much larger proportion of individuals categorized as without CHD

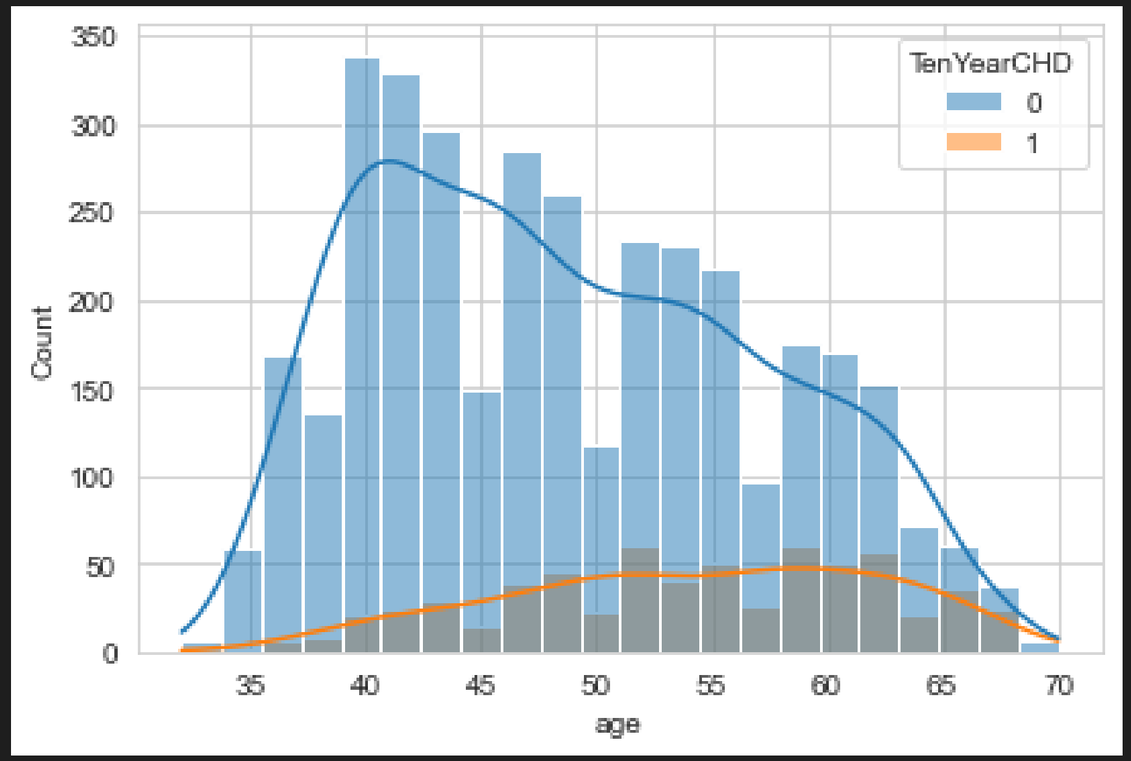
-(TenYearCHD = 0) compared to those with CHD (TenYearCHD = 1).

The analysis reveals distinct age distributions between individuals without and those with heart disease:

1. Individuals without Heart Disease (TenYearCHD = 0):
- Predominantly concentrated in the age group of 40 to 45 years.

2. Individuals with Heart Disease (TenYearCHD = 1):
- Primarily observed in the age range of 55 to 60 years.

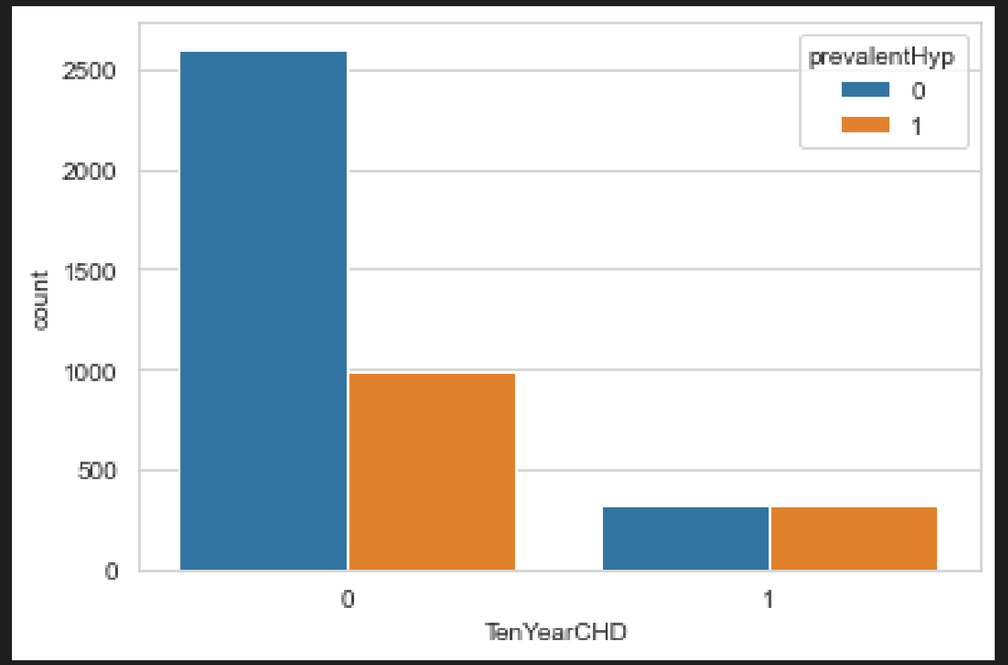
```
<AxesSubplot:xlabel='age', ylabel='Count'>
```



The analysis indicates a significant association between hypertension and heart disease within the dataset. Individuals without hypertension typically do not exhibit heart disease, while those diagnosed with heart disease tend to have a prevalence of hypertension.

```
sns.countplot(x='TenYearCHD',hue='prevalentHyp',data=data)
```

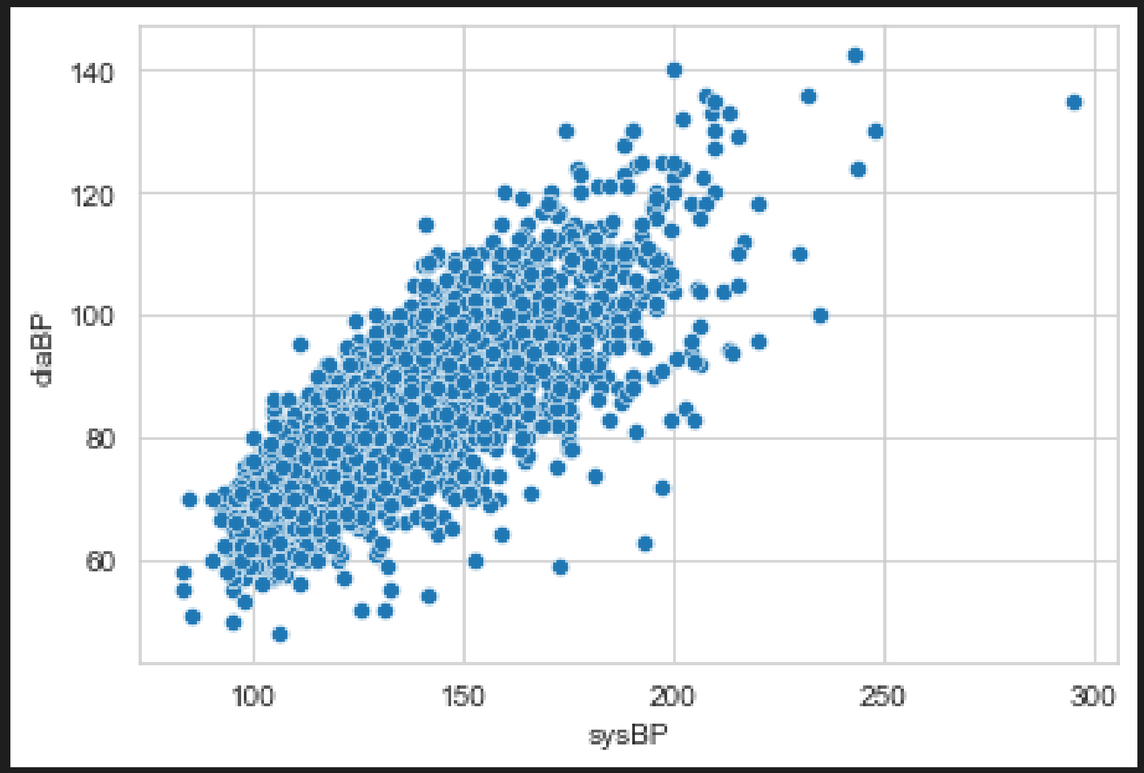
```
<AxesSubplot:xlabel='TenYearCHD', ylabel='count'>
```



The analysis shows a strong positive correlation between systolic blood pressure (SBP) and diastolic blood pressure (DBP). This indicates that as SBP increases, DBP tends to increase as well, and vice versa, highlighting their interdependence in cardiovascular health assessments.

```
sns.scatterplot(x='sysBP',y='diaBP',data=data)
```

```
<AxesSubplot:xlabel='sysBP', ylabel='diaBP'>
```

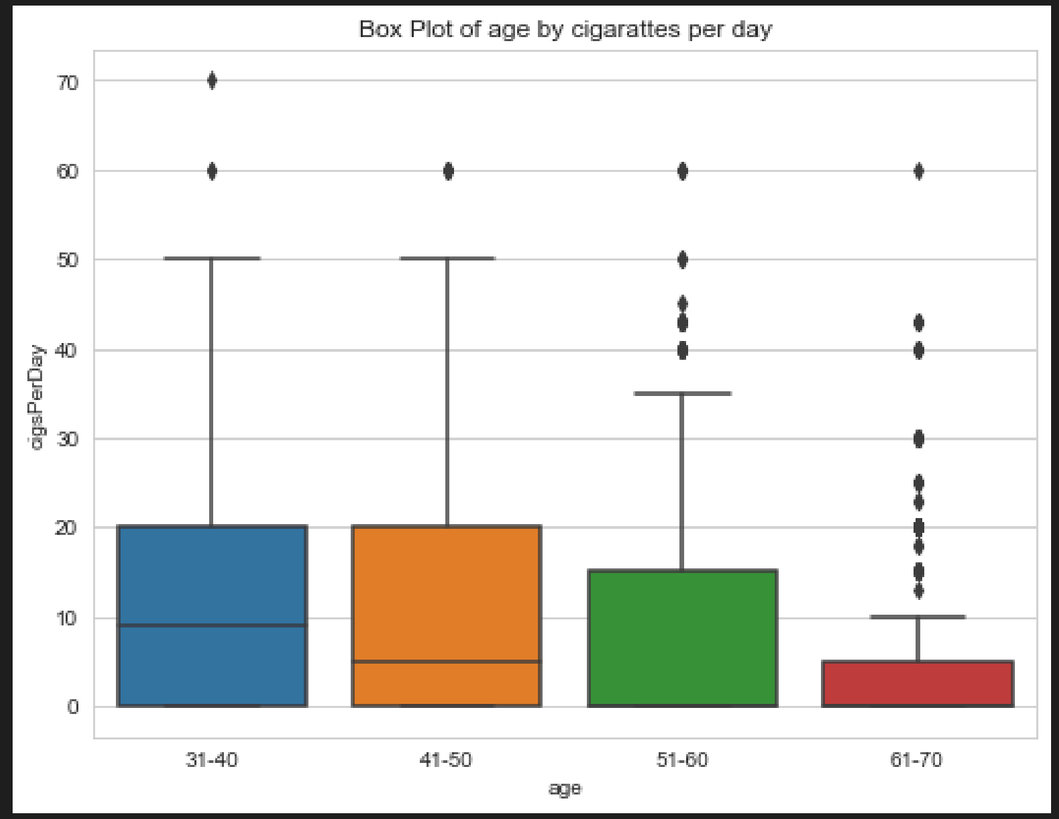


```
# Define the bin edges
bin_edges = [ 30, 40, 50, 60, 70]  # Example bin edges, adjust as needed

# Define the bin labels
bin_labels = ['31-40', '41-50', '51-60', '61-70']  # Example bin labels

# Create a new column with age bins
data['Age_Bins'] = pd.cut(data['age'], bins=bin_edges, labels=bin_labels)

# Display the DataFrame with age bins
```

I have discretized the 'age' variable into age bins to render it categorical, thereby facilitating enhanced data visualization and insights extraction.

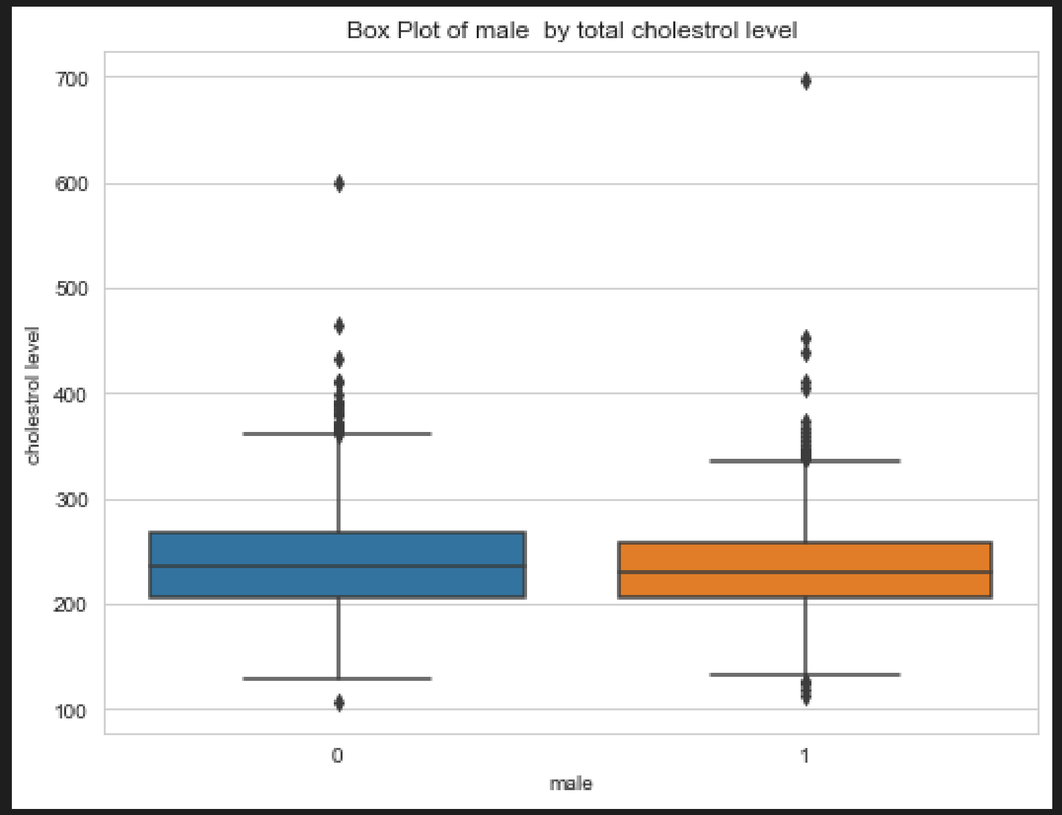
Individuals in the older age group of 61 to 70 tend to smoke a minimum of 5 cigarettes per day, which may elevate the risk of coronary heart disease. Conversely, younger individuals aged 31 to 40 are observed to smoke an average of 30 cigarettes per day, while those in the age range of 51 to 60 tend to smoke approximately 15 cigarettes per day.

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Age_Bins', y='cigsPerDay', data=data)
plt.title('Box Plot of age by cigarattes per day')
plt.xlabel('age')
plt.ylabel('cigsPerDay')
plt.show()
```



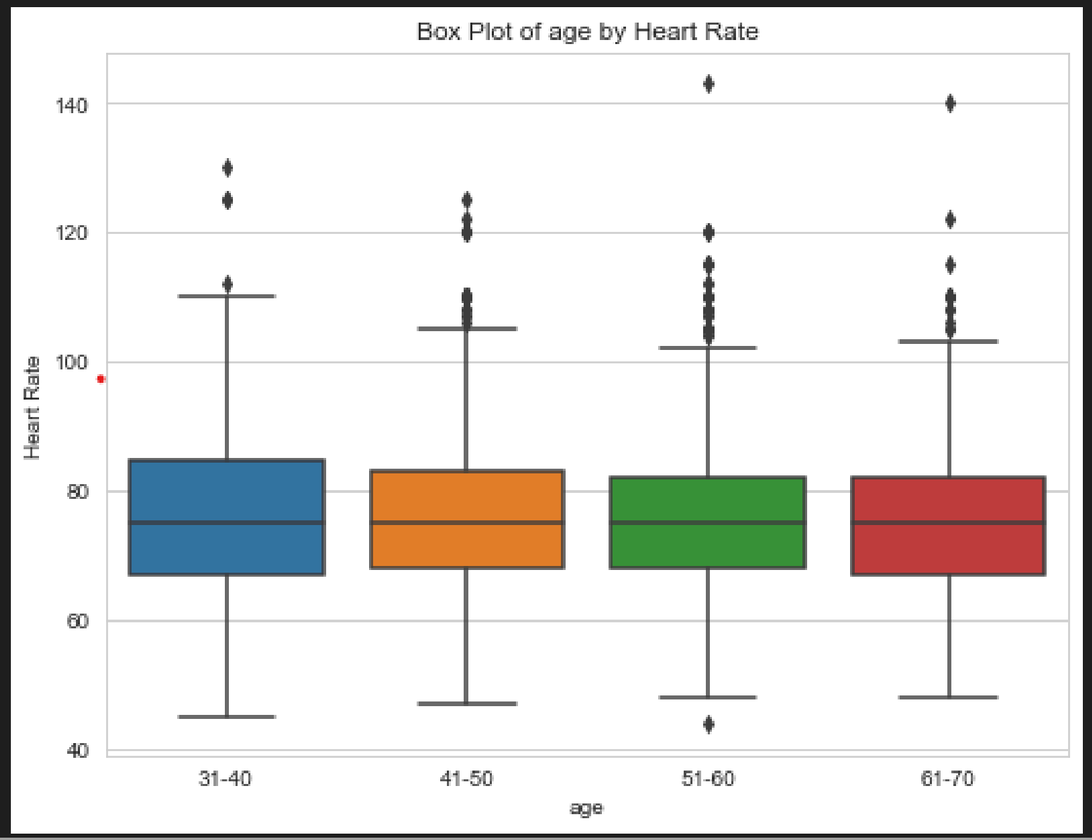
Both females and males exhibit an average cholesterol level of 250. However, a significant proportion of individuals, particularly females, display cholesterol levels of 370 or above. This pattern suggests a notable prevalence of elevated cholesterol levels across genders within the dataset.

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='male', y='totChol', data=data)
plt.title('Box Plot of male  by total cholestrol level')
plt.xlabel('male')
plt.ylabel('cholestrol level')
plt.show()
```



Across all age groups, the average heart rate remains consistent at 70 beats per minute. However, outliers are observed, indicating instances where heart rates exceed 100 beats per minute in certain individuals

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Age_Bins', y='heartRate', data=data)
plt.title('Box Plot of age by Heart Rate')
plt.xlabel('age')
plt.ylabel('Heart Rate')
plt.show()
```



CONCLUSIONS

- Middle-aged individuals (40-55 years) show a higher propensity for cardiac events, with heart disease more prevalent in those aged 55-60 years.
- The population's average cholesterol level (236.7 mg/dL) exceeds recommended levels, posing a significant cardiovascular risk.
- Severe hypertension (systolic BP of 295 mmHg) is strongly associated with heart attack risk.
- Smoking is a prevalent risk factor, with an average of 9 cigarettes per day, increasing susceptibility to heart attacks and mortality.
- Males have a higher risk of heart attacks, while females exhibit better survival rates and slightly lower average ages.
- Both genders have elevated average cholesterol levels, with females showing a notable prevalence of extreme values (370 mg/dL and above).

- Hypertension is significantly associated with heart disease, while diabetes prevalence is slightly higher in males but remains low overall.
- There is a strong positive correlation between systolic and diastolic blood pressure, indicating interdependence.
- Most individuals have healthy BMI levels and a consistent heart rate of 70 BPM, though outliers suggest potential cardiovascular risks.

THANK YOU